University of Southern Denmark
IMADA
Arthur Zimek
Jonatan Møller Gøttcke, Jonas Herskind Sejr

DM566/DM868/DM870: Data Mining and Machine Learning

Spring term 2019

Exercise 6: Clustering: k-means and Silhouette

Exercise 6-1 k-means 1-dimensional Example

Given are the following 1-dimensional points: $\{2, 3, 4, 10, 11, 12, 20, 25, 30\}$. We set $k = 3$ and choose as initial means: $\mu_1 = 2$, $\mu_2 = 4$, and $\mu_3 = 6$.

Compute the new clusters after each iteration of k-means (Lloyd/Forgy) until convergence.

Exercise 6-2 Classification vs. Clustering

Which of the following problems are classification problems, which are clustering problems?

(a) Emails in the inbox shall be sorted into spam and non-spam.

(b) Users in a database shall be grouped according to their buying patterns in past transactions.

(c) In a supermarket, products shall be placed close to each other if they are often bought together, in order to increase selling rates.

(d) Spam-mails shall be analyzed to see if there are different types of advertisements.

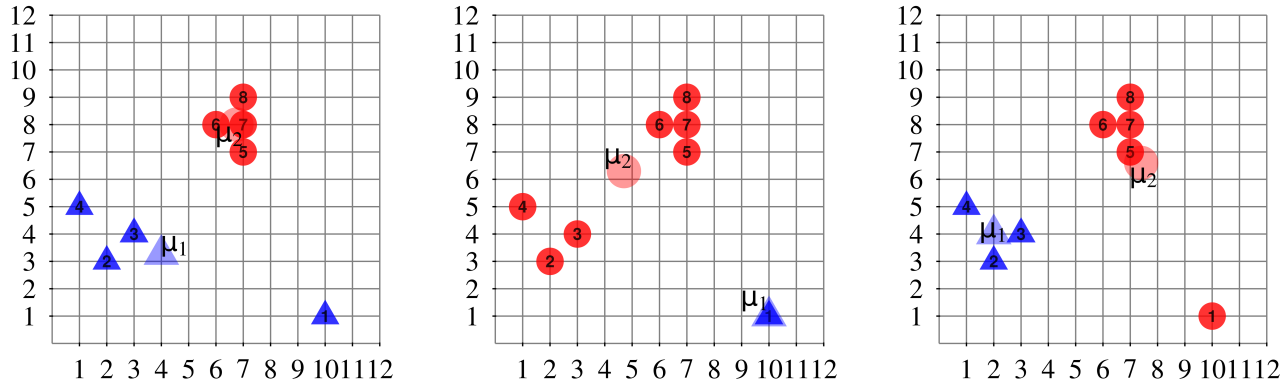
(e) Based on the DNA of some person shall be predicted of the person will suffer from diabetes within the next ten years.

(f) Data from patients with cardiac insufficiencies shall be analyzed tosee if there are groups where particular therapies work better than for others.

(g) Webpages shall be grouped into categories like “sport”, “economy”, “entertainment”.

Exercise 6-3 Silhouette Coefficient

We derived three different clustering solutions for the toy data set in the lecture:



Compute the simplified silhouette coefficient for each solution. Compare the result with the ranking by the k-means objective function (TD^2).

Exercise 6-4 Silhouette and k-means implementations in scikit-learn

Explore the code on `http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html`. To do this as lab session in class, bring your laptop with Python, NumPy, SciPy, and SciKit-Learn installed.

(a) What is the termination criterion in k-means in the scikit-learn implementation?

(b) Why can we get negative Silhouettes in this example?